(19) BUNDESREPUBLIK DEUTSCHLAND

DEUTSCHES PATENT- UND MARKENAMT

(12) **Offenlegungsschrift**

(10) **DE 197 53 454 A 1**

(51) Int. Cl.[6]: G 10 L 5/02

(21) Aktenzeichen: 197 53 454.6
(22) Anmeldetag: 2. 12. 97
(43) Offenlegungstag: 12. 11. 98

(30) Unionspriorität:
97-17615 08. 05. 97 KR

(71) Anmelder:
Electronics and Telecommunications Research Institute, Daejeon, KR

(74) Vertreter:
Betten & Resch, 80469 München

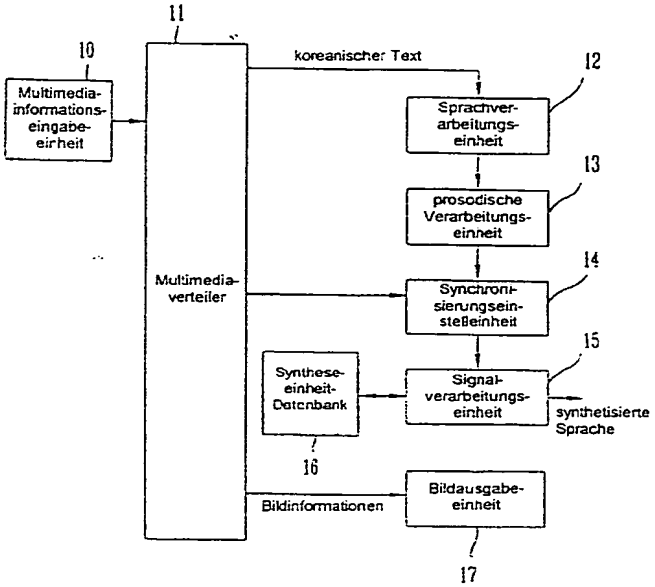
(72) Erfinder:
Lee, Jung Chul, Daejon, KR; Hahn, Min Soo, Daejon, KR; Lee, Hang Seop, Daejon, KR

Die folgenden Angaben sind den vom Anmelder eingereichten Unterlagen entnommen

Prüfungsantrag gem. § 44 PatG ist gestellt

(54) Text/Sprache-Umsetzungssystem zur Verschränkung in einer Multimediaumgebung und Verfahren zum Organisieren der Eingangsdaten für dieses System

(57) Text/Sprache-Umsetzungssystem (TTS-System) für die Verschränkung mit einer Multimediaumgebung sowie Verfahren zum Organisieren der Eingangsdaten für dieses System zum Verbessern der Natürlichkeit der synthetisierten Sprache und zum Bewerkstelligen der Synchronisierung der Multimediaumgebung mit dem TTS-System durch Definieren zusätzlicher prosodischer Informationen, Informationen, die für die Verschränkung des TTS-Systems in der Multimediaumgebung erforderlich sind, und der Schnittstelle zwischen diesen Informationen und dem TTS-System für die Verwendung bei der Erzeugung der synthetisierten Sprache. Gemäß der vorliegenden Erfindung kann ein fremdsprachiger Film in koreanischer Sprache nachsynchronisiert werden, indem die Synchronisierung der synthetisierten Sprache mit dem Film implementiert wird durch die direkte Verwendung der Textinformationen und der Lippenforminformationen, die anhand der Analyse der aktuellen Sprachdaten und der Lippenform im Film geschätzt werden, für die Erzeugung der synthetisierten Sprache. Die vorliegende Erfindung kann ferner auf verschiedene Gebiete wie z. B. einen Kommunikationsdienst, die Büroautomatisierung, die Erziehung usw. angewendet werden, indem die Synchronisierung zwischen der Bildinformation und dem TTS-System in einer Multimediaumgebung ermöglicht wird.



DE 197 53 454 A 1

DE 197 53 454 A 1

Beschreibung

Die vorliegende Erfindung bezieht sich auf ein Text/Sprache-Umsetzungssystem (im folgenden mit TTS-System bezeichnet) zum Verschränken in einer Multimediaumgebung sowie ein Verfahren zum Organisieren der Eingangsdaten für dieses System und insbesondere auf ein Text/Sprache-Umsetzungssystem (TTS-System) zur Verschränkung in einer Multimediaumgebung und ein Verfahren zum Organisieren der Eingangsdaten dieses Systems, um die Natürlichkeit der synthetisierten Sprache zu verbessern und die Synchronisierung zwischen der Multimediaumgebung und dem TTS-System zu erreichen, indem zusätzlich prosodische Informationen, die für die Verschränkung des TTS-Systems mit der Multimediaumgebung erforderlichen Informationen und eine Schnitt stelle zwischen diesen Informationen und dem TTS-System für die Verwendung bei der Erzeugung der synthetisierten Sprache definiert werden.

Der Zweck des Sprachsynthesizers ist im allgemeinen, für einen Menschen, der einen Computer benutzt, unterschiedliche Formen von Informationen zur Verfügung zu stellen. Zu diesem Zweck sollte der Sprachsynthesizer den Benutzer mit aus einem gegebenen Text synthetisierter Sprache mit hoher Qualität bedienen. Außerdem sollte der Sprachsynthesizer für die Verschränkung mit der Datenbank, die in einer Multimediaumgebung, wie z. B. mit einem Film oder einer Animation, oder mit verschiedenen Medien, die von einer Gegenseite der Konversation zur Verfügung gestellt werden, erzeugt worden ist, die synthetisierte Sprache synchron zu diesen Medien erzeugen. Die Synchronisierung des TTS-Systems mit der Multimediaumgebung ist insbesondere wichtig, um den Benutzer einen Dienst mit hoher Qualität zur Verfügung zu stellen.

Wie in Fig. 3 gezeigt, durchläuft ein typisches herkömmliches TTS-System einen aus drei Stufen bestehenden Prozeß wie folgt, bis die synthetisierte Sprache aus einem eingegebenen Text erzeugt wird.

In einem ersten Schritt setzt ein Sprachprozessor **1** den Text in eine Serie von Phonemen um, schätzt prosodische Informationen und symbolisiert diese Informationen. Das Symbol der prosodischen Information wird anhand einer Grenze der Phrasen und des Satzes, einer Position der Betonung im Wort, eines Satzmusters usw. unter Verwendung der Analyseergebnisse der Syntax geschätzt.

In einem zweiten Schritt berechnet ein prosodischer Prozessor **2** einen Wert eines prosodischen Steuerparameters anhand der symbolisierten prosodischen Informationen unter Verwendung einer Regel und einer Tabelle. Der prosodische Steuerparameter enthält die Dauer des Phonems, die Tonhöhenverlauf, den Energieverlauf und die Pauseninformation.

In einem dritten Schritt erzeugt ein Signalprozessor **3** eine synthetisierte Sprache unter Verwendung einer Synthese-Datenbank **4** und der prosodischen Steuerparameter. Mit anderen Worten bedeutet dies, daß das herkömmliche TTS-System die der Natürlichkeit und der Sprechgeschwindigkeit zugeordneten Informationen im Sprachprozessor **1** und im prosodischen Prozessor **2** nur anhand des eingegebenen Textes schätzen soll.

Ferner hat das herkömmliche TTS-System die einfache Funktion zum Ausgeben von Daten, die von der Einheit als Satz eingegeben worden sind, als synthetisierte Sprache. Um die in einer Datei gespeicherten Sätze oder die über ein Kommunikationsnetz eingegebenen Sätze der Reihe nach als synthetisierte Sprache aus zugeben, ist ein Hauptsteuerprogramm erforderlich, das die Sätze aus den eingegebenen Daten liest und diese zum Eingang eines TTS-Systems sendet. Ein solches Hauptsteuerprogramm enthält ein Verfahren zum Trennen des Textes von den eingegebenen Daten und zum ermöglichen Ausgeben der synthetisierten Sprache vom Anfang bis zum Ende, ein Verfahren zum Erzeugen der synthetisierten Sprache in Verschränkung mit einem Texteditor, ein Verfahren zum Verschränken der Sätze unter Verwendung eines Graphikschnittstelle und zum Erzeugen der synthetisierten Sprache usw., wobei jedoch die Anwendbarkeit dieser Verfahren auf Text beschränkt ist.

Derzeit haben Studien über TTS-Systeme für Landessprachen in unterschiedlichen Ländern beträchtliche Fortschritte gemacht, wobei in einigen Ländern eine gewerbliche Verwendung erreicht worden ist. Dies gilt jedoch nur für die Verwendung der Synthese der Sprache aus dem eingegebenen Text. Da es unmöglich ist, nur anhand des Textes die Informationen zu schätzen, die erforderlich sind, wenn ein Film unter Verwendung eines TTS-Systems nachsynchronisiert werden soll oder wenn die natürliche Verschränkung zwischen der synthetisierten Sprache und der Multimediaumgebung, wie z. B. bei einer Animation, implementiert werden soll, gibt es außerdem mit einer Organisation des Standes der Technik kein Verfahren zum Realisieren dieser Funktionen. Ferner liegt kein Ergebnis der Studien über die Verwendung zusätzlicher Daten zur Verbesserung der Natürlichkeit der synthetisierten Sprache und der Organisation dieser Daten vor.

Es ist daher die Aufgabe der vorliegenden Erfindung, ein Text/Sprache-Umsetzungssystem (TTS-System) zur Verschränkung in einer Multimediaumgebung sowie ein Verfahren zum Organisieren der Eingangsdaten des Systems zu schaffen, um die Natürlichkeit der synthetisierten Sprache zu verbessern und eine Synchronisierung der Multimediaumgebung mit dem TTS-System zu erreichen, indem zusätzliche prosodische Informationen, die für die Verschränkung des TTS-Systems mit der Multimediaumgebung erforderlichen Informationen sowie die Schnittstelle zwischen diesen Informationen und dem TTS-System für die Verwendung bei der Erzeugung der synthetisierten Sprache definiert werden.

Diese Aufgabe wird erfindungsgemäß gelöst durch ein Text/Sprache-Umsetzungssystem, das die im Anspruch 1 angegebenen Merkmale besitzt, sowie durch ein Verfahren zum Organisieren der Eingangsdaten eines Text/Sprache-Umsetzungssystems, das die im Anspruch 2 angegebenen Merkmale besitzt. Die abhängigen Ansprüche sind auf bevorzugte Ausführungsformen gerichtet.

Weitere Merkmale und Vorteile der vorliegenden Erfindung werden deutlich beim Lesen der folgenden Beschreibung bevorzugter Ausführungsformen, die auf die beigefügten Zeichnungen Bezug nimmt; es zeigen:

Fig. 1 eine Konstruktionsansicht eines Text/Sprache-Umsetzungssystems gemäß der vorliegenden Erfindung;

Fig. 2 eine Konstruktionsansicht einer Hardware, auf die die vorliegende Erfindung angewendet wird; und

Fig. 3 die bereits erwähnte Konstruktionsansicht eines Text/Sprach-Umsetzungssystems des Standes der Technik.

Im folgenden wird die vorliegende Erfindung anhand der bevorzugten Ausführungsform genauer beschrieben.

In Fig. 2 ist eine Konstruktionsansicht der Hardware gezeigt, auf die die vorliegende Erfindung angewendet wird. Wie in Fig. 2 gezeigt, umfaßt die Hardware eine Multimediadateneingabeeinheit **5**, eine Zentraleinheit **6**, eine Synthese-Datenbank **7**, einen Digital/Analog-(D/A)-Umsetzer **8** sowie eine Bildausgabevorrichtung **9**.

Die Multimediadateneingabeeinheit **5** empfängt Daten, die Multimediadaten wie z. B. ein Bild und einen Text umfassen, und gibt diese Daten an die Zentraleinheit **6** weiter.

Die Zentraleinheit **6** verteilt die Multimediadateneingabe der vorliegenden Erfindung, stellt die Synchronisierung ein und führt einen darin enthaltenden Algorithmus zum Erzeugen der synthetisierten Sprache aus.

Die Synthese-Datenbank **7** ist eine Datenbank, die im Algorithmus zum Erzeugen der synthetisierten Sprache verwendet wird. Diese Synthese-Datenbank **7** ist in einer Speichervorrichtung gespeichert und sendet die erforderlichen Daten zur Zentraleinheit **6**.

Der Digital/Analog-(D/A)-Umsetzer **8** setzt das synthetisierte Digitalsignal in ein Analogsignal um und gibt dieses aus.

Die Bildausgabevorrichtung **9** gibt die eingegebenen Bildinformationen auf einem Bildschirm aus.

Die Tabellen 1 und 2 sind Algorithmen, die den Zustand der organisierten Multimediaeingangsinformationen zeigen, die Text, prosodische Informationen, die Informationen für die Synchronisierung mit einem Film, die Lippenform und individuelle Eigenschaftsinformationen umfassen.

(Tabelle 1)

Syntax

TTS_Sequence() {
TTS_Sequence_Start_Code
Prosody_Enable
Video_Enable
Lip_Shape_Enable
Start_Any_Place
do{
TTS_Sentence()
}while(next_bits()==TTS_Sentence_Start_Code
}

Hierbei ist TTS_Sequence_Start_Code eine Bitkette, die hexadezimal "XXXXX" dargestellt wird und einen Beginn des TTS-Satzes bezeichnet.

TTS_Sentence_ID ist eine 10-Bit-ID und stellt eine geeignete Nummer jedes TTS-Datenstroms dar.

Language_Code stellt eine Objektsprache wie z. B. Koreanisch, Englisch, Deutsch, Japanisch, Französisch und dergleichen dar, die synthetisiert werden soll.

Prosody_Enable ist ein 1-Bit-Merker und besitzt einen Wert von "1", wenn in den organisierten Daten prosodische Daten des Orginaltons enthalten sind.

Video_Enable ist ein 1-Bit-Merker und besitzt einen Wert von "1", wenn ein TTS-System mit einem Film verschränkt ist.

Lip_Shape_Enable ist ein 1-Bit-Merker und besitzt einen Wert von "1", wenn in den organisierten Daten Lippenformdaten enthalten sind.

Trick_Mode_Enable ist ein 1-Bit-Merker und besitzt einen Wert von "1", wenn die Daten so organisiert sind, daß sie einen Trickmodus unterstützen, wie z. B. Stopp, Neustart, Vorwärts und Rückwärts.

(Tabelle 2)

Syntax

```
TTS_Sentence(){
TTS_Sentence_Start_Code
Silence
if(Silence){
        Silence_Duration
        }
else{
        Gender
        Age
        if(!Video_Enable){
                Speech_Rate
                }
        Length_of_Text
        TTS_Text
        Position_in_Sentence
        if(Prosody_Enable){
                Number_of_phonemes
                Dur_Enable
                F0_Enable
                Energy_Enable
                for(j=0 ; j<Number_of_phonemes ; j++){
                        Symbol_each_phoneme
                        Dur_each_phoneme
                        F0_contour_each_phoneme
                        Energy_contour_each_phoneme
                }
        if(Video_Enable){
                Sentence_Duration
                Position_in_Sentence
                offset
                }
        if(Lip_Shape_Enable){
                Number_of_Lip_Event
                for(j=0 ; j<Number_of_Lip_Event ; j++){
                        Lip_in_Sentence
                        Lip_Shape
                        }
                }
        }
}
```

Hierbei ist TTS_Sentence_Start_Code eine Bitkette, die hexadezimal "XXXXX" dargestellt wird und einen Beginn eines TTS-Satzes bezeichnet. TTS_Sentence_Start_Code ist eine 10-Bit-ID und stellt eine geeignete Nummer jedes TTS-Datenstroms dar.

TTS_Sentence_ID ist eine 10-Bit-ID und stellt eine geeignete Nummer jedes TTS-Satzes dar, der im TTS-Strom vorhanden ist.

Silence wird gleich "1", wenn ein vorliegender Eingangsrahmen des 1-Bit-Merkers ein stiller Sprachabschnitt ist.

In der Stufe von Silence_Duration wird eine Zeitdauer des vorliegenden stillen Sprachabschnitts in Millisekunden dargestellt.

In der Stufe von Gender wird das Geschlecht einer synthetisierten Sprache unterschieden.

In der Stufe von Age wird ein Alter der synthetisierten Sprache unterschieden zwischen Kleinkindalter, Jugendalter, mittlerem Alter und hohem Alter.

Speak_Rate stellt eine Sprechgeschwindigkeit der synthetisierten Sprache dar.

In der Stufe von Length of Text wird eine Länge des eingegebenen Textsatzes durch ein Byte dargestellt.

In der Stufe von TTS_Text wird ein Satztext mit optionaler Länge dargestellt.

Dur_Enable ist ein 1-Bit-Merker und wird gleich "1", wenn in den organisierten Daten eine Zeitdauerinformation enthalten ist.

FO_Contour_Enable ist ein 1-Bit-Merker und wird gleich "1", wenn in den organisierten Daten eine Tonhöheninformation für jedes Phonem enthalten ist.

Energy_Contour_Enable ist ein 1-Bit-Merker und wird gleich "1", wenn in den organisierten Daten eine Energieinformation für jedes Phonem enthalten ist.

In der Stufe von Number_of_phonemes, ist die Anzahl der Phoneme dargestellt, die zum Synthetisieren eines Satzes benötigt werden.

In der Stufe von Symbol_each_phoneme ist ein Symbol wie z. B. IPA dargestellt, das das jeweilige Phonem repräsentiert.

Dur_each_phoneme stellt eine Zeitdauer des Phonems dar.

In der Stufe von FO_contour_each_phoneme wird ein Tonhöhenmuster des Phonems mittels eines Tonhöhenwerts des Anfangspunkts, des Mittelpunkts und des Endpunkts des Phonems dargestellt.

In der Stufe von Energy_Contur_each_phoneme wird das Energiemuster des Phonems dargestellt, wobei ein Energie-

wert des Anfangspunkts, des Mittelpunkts und des Endpunkts des Phonems in Dezibel (dB) dargestellt wird.

Sentence_Duration stellt eine Gesamtzeitdauer der synthetisierten Sprache des Satzes dar.

Position_in_Sentence stellt eine Position des vorliegenden Rahmens im Satz dar.

In der Stufe von Offset wird dann, wenn die synthetisierte Sprache mit einem Film verschränkt ist und ein Anfangspunkt des Satzes in der Bildgruppe GOP (Group Of Pictures) liegt, eine Verzögerungszeit dargestellt, die vom Anfangspunkt der GOP zum Anfangspunkt des Satzes verstreicht.

Number_of_Lip_Event stellt die Anzahl der Änderungspunkte der Lippenform im Satz dar.

Lip_Shape stellt eine Lippenform an einem Lippenformänderungspunkt des Satzes dar.

Textinformationen enthalten einen Klassifizierungscode für eine verwendete Sprache und einen Satztext. Prosodische Informationen enthalten die Anzahl der Phoneme im Satz, Phonemstrominformationen, die Dauer jedes Phonems, das Tonhöhenmuster des Phonems sowie das Energiemuster des Phonems und werden zum Verbessern der Natürlichkeit der synthetisierten Sprache verwendet. Die Synchronisierungsinformationen des Films und der synthetisierten Sprache können als das Nachsynchronisierungskonzept betrachtet werden, wobei die Synchronisierung auf drei Wegen erreicht werden kann.

Erstens mit einem Verfahren zum Synchronisieren des Films mit der synthetisierten Sprache durch die Satzeinheit, mit der die Dauer der synthetisierten Sprache unter Verwendung der Informationen über die Anfangspunkte der Sätze, die jeweilige Dauer der Sätze und die Verzögerungszeiten der Anfangspunkte der Sätze eingestellt wird. Die Anfangspunkte der jeweiligen Sätze zeigen die Stellen der Szenen an, an denen die Ausgabe der synthetisierten Sprache für den jeweiligen Satz innerhalb des Films eingeleitet wird. Die jeweilige Dauer der Sätze gibt die Anzahl der Bilder an, die die synthetisierte Sprache für den jeweiligen Satz andauert. Außerdem sollte der Film des MPEG-2- und MPEG-4-Bildkompressionstyps, bei dem das Group-Of-Picture-(GOP)-Konzept verwendet wird, nicht in einer beliebigen Szene, sondern an einem Szenenbeginn innerhalb der Gruppe der Bilder für die Reproduktion beginnen. Somit ist die Verzögerungszeit des Anfangspunkts die zum Synchronisieren der Gruppe der Bilder und dem TTS-System benötigte Information und gibt eine Verzögerungszeit zwischen der beginnenden Szene und einem Sprachanfangspunkt an. Dieses Verfahren ist leicht zu realisieren und minimiert den zusätzlichen Aufwand, wobei es jedoch schwierig ist, eine natürliche Synchronisierung zu erreichen.

Zweitens mit einem Verfahren, mit dem die Anfangspunktinformationen, die Endpunktinformationen und die Phoneminformationen für jedes Phonem innerhalb eines Intervalls, das einem Sprachsignal im Film zugeordnet ist, markiert werden, wobei diese Informationen verwendet werden, um die synthetisierte Sprache zu erzeugen. Dieses Verfahren hat den Vorteil, daß der Grad der Genauigkeit hoch ist, da die Synchronisierung des Films mit der synthetisierten Sprache durch die Phonemeinheit erreicht werden kann, hat jedoch den Nachteil, daß ein zusätzlicher Aufwand erforderlich ist, um die Zeitdauerinformationen mit der Phonemeinheit innerhalb des Sprachintervalls des Films zu detektieren und aufzuzeichnen.

Drittens mit einem Verfahren zum Aufzeichnen der Synchronisationsinformationen auf der Grundlage des Anfangspunkts der Sprache, des Endpunkts der Sprache, der Lippenform und eines Zeitpunkts der Lippenformänderung. Die Lippenform wird quantisiert als der Abstand (Maß der Öffnung) zwischen der Oberlippe und der Unterlippe, der Abstand (Maß der Breite) zwischen den linken und rechten Punkten der Lippe und das Maß des Vorstehens der Lippe und wird als quantisiertes und normiertes Muster in Abhängigkeit vom Artikulationsort und der Artikulationsart des Phonems auf der Grundlage eines Musters mit hoher Unterscheidungsfähigkeit definiert. Dieses Verfahren ist ein Verfahren zum Steigern der Effizienz der Synchronisierung, wobei der zusätzliche Aufwand zum Erzeugen der Informationen für die Synchronisierung minimiert werden kann.

Die organisierten Multimediaeingangsinformationen, die der vorliegenden Erfindung zugeführt werden, ermöglichen einem Informationslieferanten, optional unter drei Synchronisierungsverfahren wie oben beschrieben auszuwählen und dieses zu implementieren.

Ferner werden die organisierten Multimediaeingangsinformationen zum Implementieren der Lippenanimation verwendet. Die Lippenanimation kann implementiert werden unter Verwendung des Phonemstroms, der aus dem eingegebenen Text im TTS-System und der Dauer jedes Phonems, oder unter Verwendung des Phonemstroms, der von den Eingangsinformationen verteilt wird, und der Dauer jedes Phonems, oder unter Verwendung der Informationen über die Lippenform, die in den eingegebenen Informationen enthalten sind, vorbereitet worden ist.

Die individuelle Eigenschaftsinformation erlaubt dem Benutzer, das Geschlecht, das Alter und die Sprechgeschwindigkeit der synthetisierten Sprache zu ändern. Das Geschlecht kann männlich oder weiblich sein, während das Alter in vier Stufen klassifiziert wird, z. B. 6–7 Jahre, 18 Jahre, 40 Jahre und 65 Jahre. Die Änderung der Sprechgeschwindigkeit kann zehn Stufen zwischen dem 0,7fachen und dem 1,6fachen einer Normgeschwindigkeit umfassen. Die Qualität der synthetisierten Sprache kann unter Verwendung dieser Informationen diversifiziert werden.

Fig. 1 ist eine Konstruktionsansicht des Text/Sprache-Umsetzungssystems (TTS) gemäß der vorliegenden Erfindung. Wie in **Fig.** 1 gezeigt, umfaßt das TTS-System eine Multimediainformationseingabeeinheit **10**, einen Datenverteiler für jedes Medium **11**, einen genormten Sprachprozessor **12**, einen prosodischen Prozessor **13**, eine Synchronisierungseinstellvorrichtung **14**, einen Signalprozessor **15**, eine Syntheseeinheit-Datenbank **16** sowie eine Bildausgabevorrichtung **17**.

Die Multimediaeingabeeinheit **10** ist in Form der Tabelle 1 und 2 konfiguriert und umfaßt Text, prosodische Informationen, die Informationen für die Synchronisierung mit einem Film und die Informationen über die Lippenform. Von diesen ist der Text die notwendige Information, während die anderen Informationen von einem Informationslieferanten optional als optionales Element zum Verbessern der individuellen Eigenschaft und der Natürlichkeit und zum Erreichen der Synchronisierung mit der Multimediaumgebung zur Verfügung gestellt werden können, wobei sie bei Bedarf von einem TTS-Benutzer mittels einer Zeicheneingabevorrichtung (Tastatur) oder einer Maus geändert werden können. Diese Informationen werden über das jeweilige Medium **11** zum Datenverteiler gesendet.

Der Datenverteiler empfängt über das jeweilige Medium **11** die Multimediainformationen, von denen die Bildinformationen zur Bildausgabevorrichtung **17** gesendet werden, der Text zum Sprachprozessor **12** gesendet wird und die Syn-

chronisierungsinformationen in eine Datenstruktur, die in der Synchronisierungseinstellvorrichtung **14** verwendet werden können, umgesetzt und zur Synchronisierungseinstellvorrichtung **14** gesendet werden. Wenn in den eingegebenen Multimediainformationen prosodische Informationen enthalten sind, werden diese Multimediainformationen in eine Datenstruktur umgesetzt, die der Signalprozessor **15** verwenden kann, und werden anschließend zum prosodischen Prozessor **13** und zur Synchronisierungseinstellvorrichtung **17** gesendet. Wenn in den eingegebenen Multimediainformationen individuelle Besitzinformationen enthalten sind, werden diese Multimediainformationen in eine Datenstruktur umgesetzt, die in der Syntheseeinheit-Datenbank **16** und im prosodischen Prozessor **13** innerhalb des TTS-Systems verwendet werden können, und werden anschließend zur Syntheseeinheit-Datenbank **16** und zum prosodischen Prozessor **13** gesendet.

Der Sprachprozessor **12** konvertiert den Text zu einem Phonemstrom, schätzt die prosodischen Informationen, symbolisiert diese Informationen und sendet anschließend die symbolisierten Informationen zum prosodischen Prozessor **13**. Das Symbol der prosodischen Informationen wird anhand einer Grenze der Phrase und des Satzes, einer Position der Betonung im Wort, eines Satzmusters usw. unter Verwendung des Analyseergebnisses der Syntax geschätzt.

Der prosodische Prozessor **13** empfängt das Verarbeitungsergebnis des Sprachprozessors **12** und berechnet einen Wert des prosodischen Steuerparameters, der sich von dem prosodischen Parameter unterscheidet, der in den Multimediainformationen enthalten ist. Der prosodische Steuerparameter enthält die Dauer, den Tonhöhenverlauf, den Energieverlauf, den Pausenpunkt und die Pausenlänge des Phonems. Das berechnete Ergebnis wird zur Synchronisierungseinstellvorrichtung **14** gesendet.

Die Synchronisierungseinstellvorrichtung **14** empfängt das Verarbeitungsergebnis des prosodischen Prozessors **13** und stellt für jedes Phonem die Dauer ein, um das Ergebnis mit dem Bildsignal zu synchronisieren. Die Einstellung der Dauer jedes Phonems nutzt die vom Datenverteiler über das jeweilige Medium **11** gesendete Synchronisierungsinformation. Zuerst wird jedem Phonem in Abhängigkeit vom Artikulierungsort und der Artikulierungsart des Phonems eine Lippenform zugewiesen, wobei auf der Grundlage hiervon die zugewiesene Lippenform mit der Lippenform verglichen wird, die in der Synchronisierungsinformation enthalten ist, woraufhin der Phonemstrom anhand der Anzahl der in den Synchronisierungsinformationen aufgezeichneten Lippenformen in kleine Gruppen unterteilt wird. Ferner wird die Dauer des Phonems in den kleinen Gruppen erneut unter Verwendung der Zeitdauerinformationen der Lippenform berechnet, die in der Synchronisierungsinformation enthalten ist. Die Informationen über die eingestellte Dauer werden zum Signalprozessor **15** übertragen, der das Verarbeitungsergebnis des prosodischen Prozessors **13** enthält.

Der Signalprozessor **15** empfängt die prosodische Information vom Multimediaverteiler **11** oder das Verarbeitungsergebnis der Synchronisierungseinstellvorrichtung **14**, um unter Verwendung der Syntheseeinheit-Datenbank **16** die synthetisierte Sprache zu erzeugen und auszugeben.

Die Syntheseeinheit-Datenbank **16** empfängt die individuelle Besitzinformation vom Multimediaverteiler **11**, wählt die zum Geschlecht und zum Alter passenden Syntheseeinheiten aus und sendet anschließend die für die Synthese benötigten Daten zum Signalprozessor **15** als Antwort auf eine Anfrage vom Signalprozessor **15**.

Wie aus der obigen Beschreibung deutlich wird, können die individuellen Eigenschaften der synthetisierten Sprache gemäß der vorliegenden Erfindung verwirklicht werden, wobei die Natürlichkeit der synthetisierten Sprache verbessert werden kann durch Organisieren der individuellen Eigenschaften und der prosodischen Informationen, die durch die Analyse der aktuellen Sprachdaten geschätzt werden, zusammen mit den Textinformationen als mehrstufige Informationen. Ferner kann ein fremdsprachiger Film in koreanischer Sprache nachsynchronisiert werden, indem die Synchronisierung der synthetisierten Sprache mit dem Film implementiert wird durch die direkte Verwendung der Textinformationen und der Lippenforminformationen, die anhand der Analyse der aktuellen Sprachdaten geschätzt werden, und der Lippenform im Film zur Herstellung der synthetisierten Sprache. Die vorliegende Erfindung kann ferner auf verschiedene Gebiete wie z. B. einem Kommunikationsdienst, der Büroautomatisierung, der Erziehung usw. angewendet werden, indem die Synchronisierung zwischen der Bildinformation und dem TTS-System in einer Multimediaumgebung ermöglicht wird.

Obwohl die vorliegende Erfindung und ihre Vorteile genau beschrieben worden sind, ist klar, daß verschiedene Änderungen, Ersetzungen und Abwandlungen daran vorgenommen werden können, ohne vom Geist und vom Umfang der Erfindung, wie sie durch die beigefügten Ansprüche definiert ist, abzuweichen.

Die beigefügten Ansprüche sollen daher alle solchen Anwendungen, Abwandlungen und Ausführungsformen innerhalb des Umfangs der Erfindung abdecken.

Patentansprüche

1. Text/Sprache-Umsetzungssystem (TTS-System) für die Verschränkung in einer Multimediaumgebung, **gekennzeichnet durch**

eine Multimediainformationseingabeeinheit (**10**) zum Organisieren von Text, prosodischen Informationen, Informationen über die Synchronisierung mit einem Film, der Lippenform und der Informationen wie z. B. der individuellen Eigenschaft;

einen Datenverteiler (**11**) zum Verteilen der Informationen der Multimediainformationseingabeeinheit (**5**) auf die Informationen für die jeweiligen Medien;

einen Sprachprozessor (**12**) zum Umsetzen des vom Datenverteiler (**11**) über das jeweilige Medium verteilten Textes in einen Phonemstrom, zum Schätzen der prosodischen Informationen und zum Symbolisieren der Informationen;

einen prosodischen Prozessor (**13**) zum Berechnen eines Werts des prosodischen Steuerparameters aus der symbolisierten prosodischen Information unter Verwendung einer Regel und einer Tabelle;

eine Synchronisierungseinstellvorrichtung (**14**) zum Einstellen der Dauer des Phonems unter Verwendung der vom Datenverteiler (**11**) über das jeweilige Medium verteilte Synchronisierungsinformation;

einen Signalprozessor (**15**) zum Erzeugen einer synthetisierten Sprache unter Verwendung des prosodischen Steu-

erparameters und der Daten in einer Syntheseeinheit-Datenbank (16); und
eine Bildausgabevorrichtung (17) zum Ausgeben der vom Datenverteiler (11) über das jeweilige Medium vereilten Bildinformationen auf einem Bildschirm.

2. Verfahren zum Organisieren der Eingangsdaten eines Text/Sprache-Umsetzungssystems (TTS-System) für die Verschränkung in einer Multimediaumgebung, gekennzeichnet durch die Schritte:
Klassifizieren der Multimediaeingangsinformationen, die zur Verbesserung der Natürlichkeit der synthetisierten Sprache und zur Implementierung der Synchronisierung der Multimediaumgebung mit dem TTS-System organisiert sind, in Text, prosodische Informationen, Informationen für die Synchronisierung mit einem Film, die Lippenform und die individuellen Eigenschaftsinformationen in einer Multimediainformationseingabeeinheit (10);
Verteilen der in der Multimediainformationseingabeeinheit (10) klassifizierten Informationen in einem Datenverteiler (11) auf die jeweiligen Medien auf der Grundlage entsprechender Informationen;
Umsetzen des im Datenverteiler (11) auf die jeweiligen Medien verteilten Textes in einen Phonemstrom, Schätzen der prosodischen Informationen und Symbolisieren der Informationen in einem Sprachprozessor (12);
Berechnen eines Werts des prosodischen Steuerparameters, die sich von dem prosodischen Steuerparameter unterscheidet, der in den Multimediainformationen enthalten ist, in einem prosodischen Prozessor (13);
Einstellen der Dauer jedes Phonems in einer Synchronisierungseinstellvorrichtung (14), so daß das Verarbeitungsergebnis im prosodischen Prozessor (13) mit einem Bildsignal gemäß der Eingabe der Synchronisierungsinformationen synchronisiert werden kann;
Erzeugen der synchronisierten Sprache in einem Signalprozessor (15) unter Verwendung der prosodischen Informationen vom Datenverteiler (11), des Verarbeitungsergebnisses in der Synchronisierungseinstellvorrichtung (14) und einer Syntheseeinheit-Datenbank (16); und
Ausgeben der Bildinformationen, die vom Datenvereiler (11) über die jeweiligen Medien verteilt worden sind, auf einem Bildschirm in einer Bildausgabevorrichtung (17).

3. Verfahren nach Anspruch 2, dadurch gekennzeichnet, daß die organisierten Multimediainformationen Textinformationen, prosodische Informationen, Informationen für die Synchronisierung mit einem Film, Lippenforminformationen und Individualitätsinformationen enthalten.

4. Verfahren nach Anspruch 3, dadurch gekennzeichnet, daß die prosodischen Informationen die Anzahl der Phoneme, die Phonemstrominformationen, die Zeitdauer jedes Phonems, das Tonhöhenmuster des Phonems und das Energiemuster des Phonems umfassen.

5. Verfahren nach Anspruch 4, dadurch gekennzeichnet, daß die Dauer des Phonems einen Wert der Tonhöhe am Anfangspunkt, am Mittelpunkt und am Endpunkt innerhalb des Phonems angibt.

6. Verfahren nach Anspruch 4, dadurch gekennzeichnet, daß das Energiemuster des Phonems einen Energiewert in Dezibel am Anfangspunkt, am Mittelpunkt und am Endpunkt innerhalb des Phonems angibt.

7. Verfahren nach Anspruch 2, dadurch gekennzeichnet, daß die Synchronisierungsinformationen einen Text, eine Lippenform, eine Positionsinformation innerhalb eines Films und die Zeitdauerinformation umfassen.

8. Verfahren nach Anspruch 2, dadurch gekennzeichnet, daß die Synchronisierungsinformationen einen Anfangspunkt, eine Dauer und eine Verzögerungszeitinformation des Startpunkts umfassen, wobei die Dauer jedes Phonems durch diese Synchronisierungsinformationen gesteuert wird.

9. Verfahren nach Anspruch 2, dadurch gekennzeichnet, daß die Synchronisierungsinformationen eine Dauer des Anfangspunktes eines Satzes und eine Zeitdauerinformation des Startpunkts umfassen, wobei die Dauer jedes Phonems durch die vorhergesagte Lippenform unter Berücksichtigung einer Artikulationsart des Phonems und die Artikulationssteuerung gesteuert wird, wobei die Lippenform innerhalb der Synchronisierungs- und Zeitdauerinformationen die Synchronisierungsinformationen bilden.

10. Verfahren nach Anspruch 2, dadurch gekennzeichnet, daß die synchronisierte Sprache anhand einer Information über den Anfangspunkt und den Endpunkt jedes Phonems, das dem Sprachsignal zugeordnet ist, und anhand einer Information des Phonems erzeugt wird.

11. Verfahren nach Anspruch 2, dadurch gekennzeichnet, daß die synchronisierte Sprache anhand einer Quantisierung des Abstandes (Maß der Öffnung) zwischen der Oberlippe und der Unterlippe, eines Abstandes (Maß der Breite) zwischen den linken und rechten Endpunkten einer Lippe und eines Maßes des Vorstehens einer Lippe erzeugt wird, wobei die Lippenform ein quantisiertes und normiertes Muster ist, das vom Artikulationsort und der Artikulationsart des Phonems auf der Grundlage des Musters mit starken Unterscheidungsmerkmalen ist.

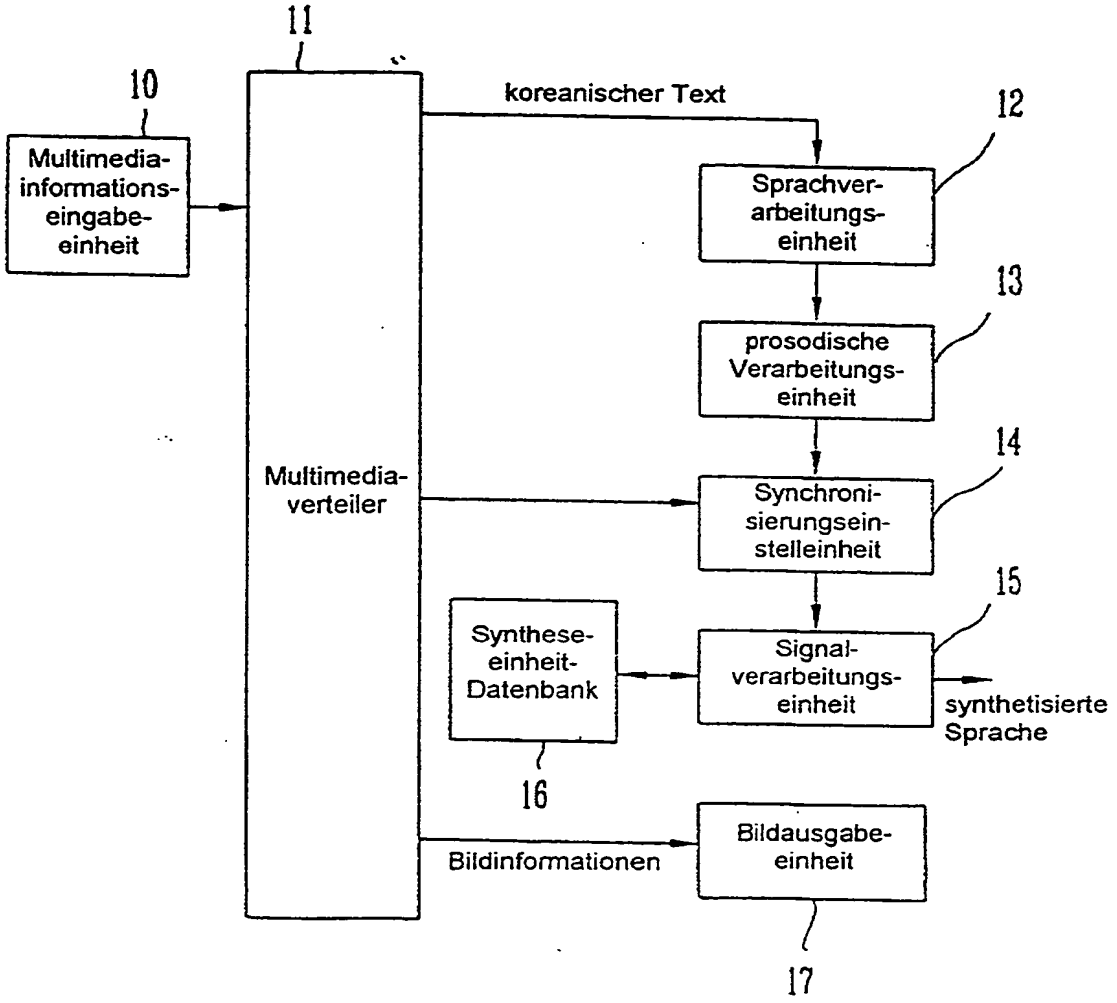
12. Verfahren nach Anspruch 2, dadurch gekennzeichnet, daß
das Sendeverfahren der Multimediainformationen die Schritte umfaßt:
Umsetzen einer in den Multimediainformationen vorhandenen prosodischen Information in eine Datenstruktur, die im Signalprozessor (12) verwendet werden kann;
Senden der umgesetzten prosodischen Informationen um prosodischen Prozessor (13) und zur Synchronisierungseinstellvorrichtung (14);
Umsetzen der vom prosodischen Prozessor (13) und von der Synchronisierungseinstellvorrichtung (14) ausgegebenen prosodischen Informationen in eine Datenstruktur, die in der Syntheseeinheit-Datenbank (16) und im prosodischen Prozessor (13) innerhalb des TTS-Systems verwendet werden kann, wenn die prosodischen Informationen in den Multimediaeingangsinformationen enthalten sind; und
Senden der Informationen zur Syntheseeinheit-Datenbank (16) und zum prosodischen Prozessor (13), wenn die individuellen Eigenschaftsinformationen in den Multimediaeingangsinformationen enthalten sind.

Hierzu 2 Seite(n) Zeichnungen

- Leerseite -

Nummer: DE 197 53 454 A1
Int. Cl.[6]: G 10 L 5/02
Offenlegungstag: 12. November 1998

FIG. 1



Nummer: DE 197 53 454 A1
Int. Cl.[6]: G 10 L 5/02
Offenlegungstag: 12. November 1998

FIG. 3

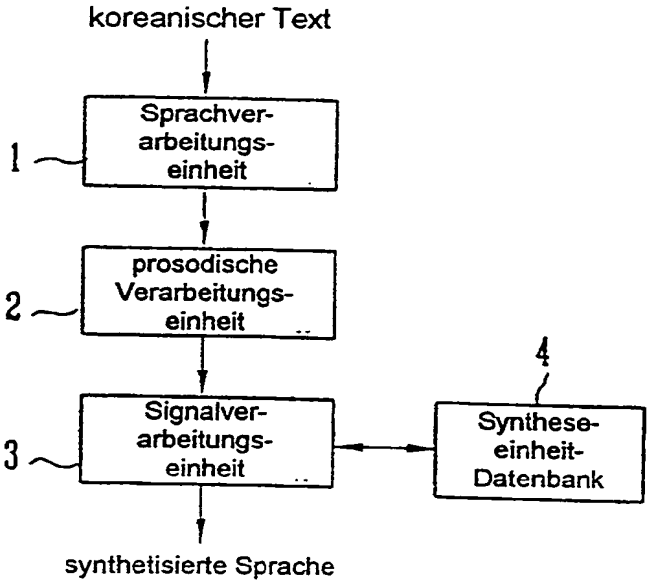


FIG. 2